

## Essay: Data Wrangling Efforts in the WeRateDogs Twitter Archive Project

Data wrangling is crucial in data analysis, ensuring data is accurate, consistent, and structured appropriately for meaningful insights. The *"wrangle\_act.ipynb"* notebook demonstrates a systematic approach to data wrangling using the popular WeRateDogs Twitter dataset. The primary goal of this project was to collect, assess, clean, and prepare various data sources related to dog ratings on Twitter, enabling compelling exploration and analysis.

The wrangling process began with **data gathering** from three distinct sources. First, a CSV file containing the WeRateDogs Twitter archive was directly downloaded. Second, a TSV file with tweet image predictions generated by a neural network was retrieved using a programmatic request. Third, Twitter's API collected additional tweet metadata using the Tweepy library. Each source required a different access method, showcasing various data acquisition techniques in real-world projects.

Once gathered, the next step was **data assessment**. We conducted the evaluation visually and programmatically to identify data quality and structure issues. Several common data quality issues emerged, such as missing values, inconsistent naming conventions for dog breeds, incorrect data types, and duplicated or irrelevant columns. Structural problems were also observed, such as the need to combine multiple columns that referred to a single variable (e.g., different stages of a dog).

Following the assessment, the **data cleaning** phase was initiated. This involved resolving each identified issue through careful manipulation using the pandas library. Invalid or missing ratings were corrected or removed, dog names that were not real (e.g., "a" or "an") were replaced with None, and data types were converted to appropriate formats (e.g., timestamps and numerics). Additionally, image prediction results were filtered to retain only the most confident predictions, and multiple data sources were merged into a single master dataset for unified analysis.

Throughout the wrangling process, the notebook maintained a high level of organization and documentation, with clear explanations and logical code structuring. Visual checks and assertions were often used to verify the success of cleaning steps, ensuring data integrity before moving to analysis.

In summary, this project demonstrates the importance and complexity of data wrangling. The efforts involved gathering diverse data sources, systematically identifying and addressing data issues, and combining them into a clean, structured dataset ready for exploration. Such meticulous preparation is essential to derive accurate, insightful, and reliable conclusions in any data analysis endeavor.